Multilevel models ans small area estimation in the context of Vietnam living standards surveys

Phong Nguyen, Dominique Haughton, Irene Hudson, John Boland

MULTILEVEL MODELS AND SMALL AREA ESTIMATION IN THE CONTEXT OF VIETNAM LIVING STANDARDS SURVEYS

Phong Nguyen, Dominique Haughton, Irene Hudson and John Boland

General Statistics Office, Hanoi, Vietnam, Bentley University, Waltham, MA, USA and Toulouse School of Economics, Toulouse, France, University of South Australia, University of South Australia, Adelaide, Australia

Abstract: This talk discusses a methodology to obtain small area estimates in the context of the Vietnam Living Standards Surveys. First we briefly introduce these surveys. Second, we recall main concepts in small area estimation, including the use of auxiliary data, and contrast simple with regression small area models. We discuss random effects in small area regression models, and, in the third part of the talk, present our proposed multilevel model for small area estimation at the commune level in Vietnam, to our knowledge the first such model built with Vietnam living standards data. Our model for estimating the commune-level mean (log of) household expenditure per capita relies on independent variables available both in the 1999 Census and in the VHLSS of 2002 and follows ideas given in work by Moura (1994, 1999); we mention how to measure the accuracy of our model.

Résumé: L'exposé présente une méthode pour obtenir des estimateurs pour petites régions dans le contexte des Enquêtes sur le Niveau de Vie au Vietnam. On introduit brièvement ces enquêtes, puis on rappelle les concepts principaux en estimation pour petites régions, notamment l'utilisation de données auxiliaires, et on contraste les modèles simples avec ceux de régression. On traite les effets aléatoires dans ces modèles et on propose un modèle multi-niveaux pour une estimation au niveau de la commune au Vietnam, à notre connaissance le premier modèle de ce type construit à partir de données sur le niveau de vie au Vietnam. Notre modèle pour la moyenne au niveau communal du logarithme des dépenses familiales par personne utilise des variables indépendantes disponibles par le Recensement de 1999 et l'enquête aux ménages de 2002 et suit des idées exposées par Moura (1994, 1999); on discute comment mesurer la précision du modèle.

VIETNAM LIVING STANDARDS SURVEYS

The model used in this paper relies on data from the Viet Nam Household Living Standards Survey (VHLSS) of 2002, and we refer to past work which relies on the Viet Nam Living Standards Surveys (VLSS) of 1993 and 1998. In this section we describe a few relevant features of the surveys, and the context in which the VLSS, and then the VHLSS program were established under the auspices of the General Statistics Office (GSO 1999) in Viet Nam. Further details are available from Nguyen Phong and Haughton (2006).

BRIEF INTRODUCTION TO SMALL AREA ESTIMATION

This paper studies whether some communes, districts and provinces had more 'effective' influence than others in promoting households' living standards, taking account of variations in the characteristics of households. To this end, we use multilevel modeling and apply this methodology to small area estimation.

Small area estimation is widely used in a number of national statistics offices over the world.

References are many, but for the purposes of this paper a very useful reference is the Small Area Estimation manual by the Australian Bureau of Statistics (ABS 2005).

Small area estimation methods are often divided into two main types of methods: "simple small area methods", such as for example direct estimation (where small area estimators are obtained directly from survey data) which typically yields an unbiased estimator but with a large standard error because of small sample sizes), and methods such as broad area ratio estimators (ABS 2005). In this paper, we focus attention on small area methods which rely on a regression model. In many applications a regression model is used with independent variables available for the entire population (such as via a census), and the model is applied to obtain estimates of for example the mean of the dependent variable at the small area level. When the regression model does not include any random effects that might capture local effects, the methods is often referred to as "synthetic regression models".

In this paper, we follow up on work by Moura and colleagues (1994, 1999) who began to promote the use of random effects in regression models to obtain improved small area estimators.

OUR MULTILEVEL MODEL

The model we have constructed is a four-level model for a one-year period using the 2002 Vietnam Household Living Standards Survey. The four levels include the household level *i*, commune level *j*, district level *k* and provincial level *l*.

The dependent variable is the logarithm of real per capita household expenditure. Independent variables include 21 variables (listed below) that reflect household characteristics (measured at the household level) from VHLSS 2002 and that are also available in the 1999 Vietnam Population and Housing Census (Census 1999). Our model in its most general form can be described as follows:

$$Y_{ijkl} = \beta_{0jkl} + \sum_{p} \beta_{pjkl} X_{pijkl} + \varepsilon_{ijkl} \quad (1)$$

$$\beta_{0jkl} = \gamma_{00} + \gamma_{01} Z_{jkl} + f_{0l} + \nu_{0kl} + u_{0jkl} \quad (2)$$

$$\beta_{pjkl} = \gamma_{p0} + \gamma_{p1} Z_{jkl} + f_{pl} + \nu_{pkl} + u_{pjkl} \quad (3)$$

where the Y_{ijkl} represent the values of the dependent variable at the first level (household level). This is in our case the logarithm of the real per capita expenditure of the i^{th} household (i=1, ..., n_j , level 1) in the j^{th} commune (j=1,..., m_k, level 2) of the k^{th} district (k=1,..., r_l, level 3) of the l^{th} province (l=1,...,61, level 4); the X_{pijkl} represent the values of the p^{th} explanatory variable measured for the i^{th} household in the j^{th} commune of the k^{th} district of the l^{th} province; here p =1,...,21, corresponding to the 21 variables listed below. Note that in VHLSS 2002, the value of n_j, in principle equal by design to 25 for all communes, in fact varies: 759 communes had more than 5 households (17-25) in the sample, and 2,142 communes had 3-5 households in the sample. The β_{0jkl} represent the regression intercepts (for each commune j in district k in province l), and the β_{pjkl} represent the regression coefficients (slopes) (for each commune j in district k in province l for each of the 21 independent variables, p=1,...,21). The error terms ε_{ijkl} represent the usual residual error terms assumed to have mean 0 and variance σ^2_{ijkl} typically assumed to be constant equal to a common error variance σ^2 (a property referred to as homoskedasticity). The Z_{jkl} denote the values of one independent variable, measured at the commune level j (in district k in province l); to simplify notations, we assume that we have only one such variable, but the model extends

easily to more than one such variable. The coefficients $\gamma_{00}, \gamma_{01}, \gamma_{p0}, \gamma_{p1}$ are fixed regression coefficients, and the u_{0jkl} and u_{pjkl} are random residual error terms at the commune level, assumed to have a mean of zero and to be independent from the ε_{ijkl}. In addition the u_{0jkl} and u_{pjkl} are assumed to have a constant variance. In a similar way, the v_{0kl} and v_{pkl} are random residual error terms at the district level, assumed to have a mean of zero and to be independent from the ε_{ijkl}, and a constant variance. Finally, the f_{0l} and f_{pl} are random residual error terms at the province level, assumed to have a mean of zero and to be independent from the ε_{ijkl} as well as to have a constant variance. The model is made multilevel by allowing the regression linear combination for each household to shift (higher or lower) from the overall linear combination by an amount $u_{0jkl} + v_{0kl} + f_{0l} + \varepsilon_{ijkl}$.

Our multilevel model in MLwiN output format is as follows.

$$\text{lrpcexp}_{ijkl} \sim N(XB, \Omega)$$

$$\begin{aligned}\text{lrpcexp}_{ijkl} = {} & \beta_{0ijkl}\text{cons} + \text{-0.039(0.009)female}_{ijkl} + \text{-0.441(0.011)children}_{ijkl} + \text{-0.054(0.011)elderly}_{ijkl} + \\ & \text{-0.081(0.001)hhsize}_{ijkl} + \beta_{5kl}\text{urban}_{jkl} + \text{0.074(0.007)safewater}_{ijkl} + \text{0.287(0.007)toiletflush}_{ijkl} + \\ & \text{0.168(0.011)toiletsuilabh}_{ijkl} + \text{0.190(0.006)housepermnt}_{ijkl} + \text{-0.173(0.005)housetem}_{ijkl} + \\ & \text{0.085(0.008)electricity}_{ijkl} + \text{0.182(0.005)tv}_{ijkl} + \text{4.667(0.927)agerescale}_{ijkl} + \text{-41.759(8.886)agerescale2}_{ijkl} + \\ & \text{0.073(0.009)kinh}_{ijkl} + \text{0.018(0.001)yearsedu}_{ijkl} + \text{0.107(0.013)leader}_{ijkl} + \text{0.147(0.016)h\_skilled}_{ijkl} + \\ & \text{0.058(0.013)m\_skilled}_{ijkl} + \text{-0.047(0.005)noskilled}_{ijkl} + \text{0.006(0.001)urbyearsed}_{ijkl}\end{aligned}$$

$$\beta_{0ijkl} = 7.893(0.035) + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

$$\beta_{5kl} = 0.076(0.018) + f_{5l} + v_{5kl}$$

$$\begin{bmatrix} f_{0l} \\ f_{5l} \end{bmatrix} \sim N(0, \Omega_f) : \Omega_f = \begin{bmatrix} 0.031(0.006) & \\ \text{-0.002(0.003)} & 0.008(0.003) \end{bmatrix}$$

$$\begin{bmatrix} v_{0kl} \\ v_{5kl} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.015(0.001) & \\ \text{-0.002(0.002)} & 0.006(0.003) \end{bmatrix}$$

$$\begin{bmatrix} u_{0jkl} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.012(0.001) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ijkl} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.086(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 14782.110(29530 of 29530 cases in use)

In our model, the 21 independent variables are as follows: *urban* (1=urban, 0=rural), *hhsize* (household size), *elderly* (proportion of elderly), *children* (proportion of children), *female* (proportion of females), *kinh* (ethnicity of head, 1=Kinh, 0= not Kinh), *agerescale* (rescaled age of household head =age/1000), *agerescale2* (squared rescaled age of household head), *yearsedu* (number of years of education of head), *urbyearsed* (interaction of urban and yearsedu), *leader* (leadership job, yes=1, otherwise=0), *h_skilled* (high skilled job, university and above=1, otherwise=0, *m_skilled* (medium skilled job, secondary professional and training=1, otherwise=0, *noskilled* (non-skilled nonfarm worker, yes=1, otherwise=0; reference= non-skilled farm worker), *housepermnt* (having permanent house=1), *housetem* (having temporary house=1, reference=semi-permanent), *electricity* (having electricity=1), *safewater* (having safe water source=1), *toiletflush* (having flushing toilet=1), *toiletsuilab* (having suilabh toilet=1, reference=other), *tv* (having a tv set=1, otherwise=0).

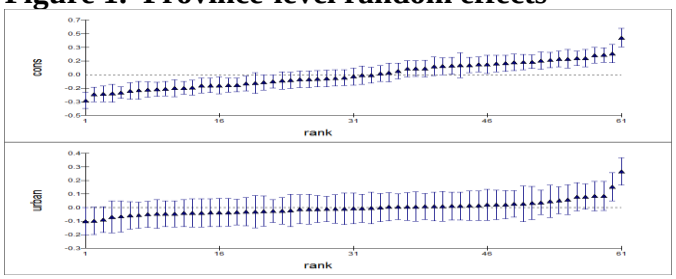
The regression coefficient for the intercept of *lrpcexp* is β_0; the regression coefficients for *lrpcexp* are β_1 to β_{21} corresponding to the 21 independent variables shown above.

These coefficients go together with standard errors in brackets; all are significant. For example, for the variable "children" (the proportion of children in each household) the coefficient is -0.441 with its standard error of 0.011, which is significant. Some coefficients include a random component, namely the intercept β_{0ijkl} and the coefficient β_{5kl} of the urban/rural variable. Since all predictors except for the urban/rural dummy variable are assigned only fixed effects, the slopes of the lines are all the same except for the urban/rural variable, but the intercepts are different for each commune, since we have assigned both fixed and random effects to the intercept. For example, for the variable "children", the fitted value of the fixed coefficient is -0.441 and its standard error is 0.011 (in bracket), as mentioned above. So for all communes the slope of the variable "children" is -0.441. The estimated fixed part of the intercept is 7.893, with a fitted standard error (in bracket) of 0.035. The intercepts for the different communes incorporate the fitted level 2 residuals u_{0jkl} which are distributed around their mean with a variance of 0.012 (standard error 0.001). The intercepts for the different districts incorporate the fitted level 3 residuals v_{0kl} which are distributed around their mean with a variance of 0.015 (standard error 0.001). The intercepts for the different provinces

incorporate the fitted level 4 residuals f_{0l} which are distributed around their mean with a variance of 0.031 (standard error 0.006). This model has random effects at the district and provincial levels which are included in the coefficient of the "urban" dummy variable. The motivation for including those random effects is to attempt to capture unexplained geographical differences in the urban/rural gap (Haughton and Nguyen 2008), known to be important in Vietnam as a source of inequality. As can be seen from the model, β_{5kl} = 0.076(0.018) + f_{5l} + v_{5kl} where the fixed effect is 0.076 with standard error 0.018, the province-level random effect (province-level residual) f_{5l} has variance 0.008 with standard error 0.003, and the district-level random effect (or district-level residual) v_{5kl} has variance 0.006 with standard error 0.003. Note that the coefficient for the "urban" dummy variable does not include a commune-level random effect, since communes are either entirely rural or entirely urban. Our model does not include variables at a higher level than the household level, for example Z_{jkl}. However, in a future model we will use some variables from the Viet Nam 2001 Agriculture Census available for all households in rural communes to build a small area model for rural areas using the VHLSS 2002.

COMMUNE-LEVEL, DISTRICT-LEVEL AND PROVINCE-LEVEL RANDOM EFFECTS

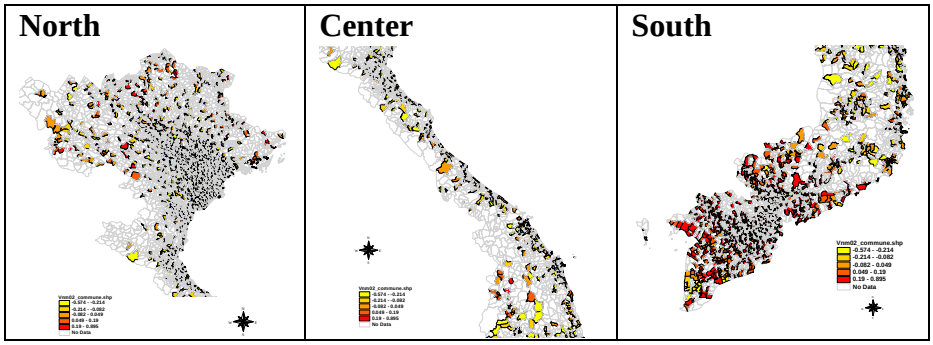
Figure 1. Province-level random effects



In order to see whether some communes, districts and provinces had a more 'effective' influence than others in promoting household living standards, we calculated commune-level random effects (u_{0jkl}), district-level random effects (v_{0kl}), and province-level random effects (f_{0l}) using MLwiN. The results of the calculation can be plotted by MLwiN as presented below for province level random effects. Note that the graphs include random effects for both the intercept ('cons') and the urban/rural dummy variable ('urban'). Note also that the random effects are plotted in increasing order, and with approximate 95% confidence intervals. Figure 1 displays 61 province level 4 random effects, one for each province. There are 37 provinces in the plot where the confidence intervals for their random effects do not overlap zero; among them there 19 provinces have negative random effects and 18 provinces have positive random effects.

We will use visual tools to display random effects. An example of a display of total intercept random effects (province plus district plus commune-level effects) in the form of a map is given below in Figure 2. Communes coloured red are communes whose location is associated with higher living standards, even once variables such as age, education and job status of the head of household are controlled for. On the other hand, communes coloured bright yellow are communes whose location is associated with lower living standards, controlling for those same variables. Such a display can be very useful to help identify communes in Vietnam that suffer challenges by their very location; these challenges could be due to isolation or particularly difficult climate conditions etc.

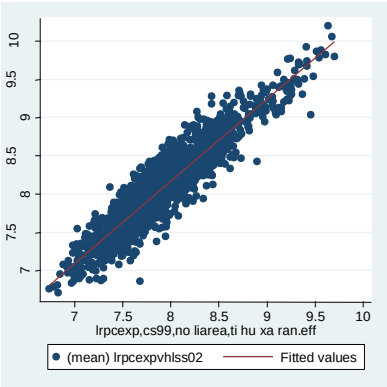
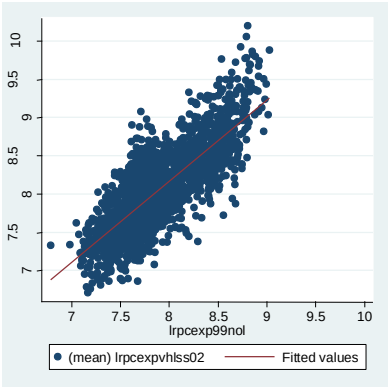
Figure 2. Total (province plus district plus commune) intercept random effects in our model for expenditure per capita



SMALL AREA (COMMUNE) ESTIMATES

Here we use our multilevel model to obtain a predictor for the population mean of small areas, following ideas suggested by Moura (1994) and Moura and Holt (1999) briefly described below. These ideas consist in plugging the commune population means of the 21 variables from the Census conducted in 1999 into the independent variables in our multilevel model to estimate the mean logarithm of real per capita expenditure for VHLSS 2002 communes.

Including random effects does improve the small area estimation, as we will see in the graph below. This graph proposed by Brown, Chambers, Heady and Heasman (Evaluation of small area estimation methods, Proceedings of Statistics Canada Symposium 2001) as a tool for checking model validity for a small area estimation shows that the estimates with random effects are closer to the least squares fit line, which is itself close to the 45-degree line:

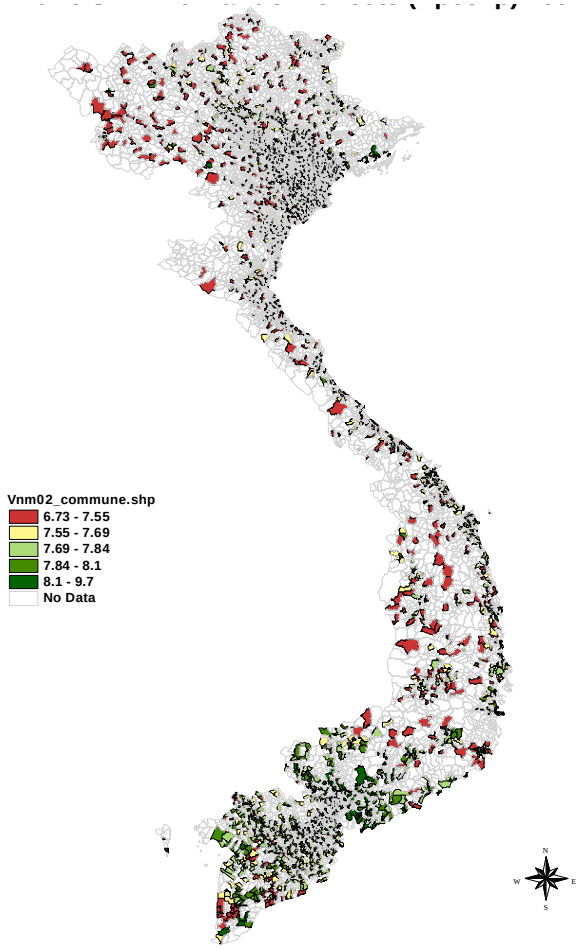
Direct estimates vs. small area estimates, *with* province, district and commune random effects	Direct estimates vs. small area estimates, *without* province, district and commune random effects
	

Note: Direct estimates are on Y axis, small area estimates are on the X axis.

The idea of this diagnostic graph is that if the small area estimates are a good representation of the "truth" – the population means, the direct survey observations should behave as a random sample from a distribution with mean equal to the population means.

A Geographical Information Systems (GIS) representation of our small area estimates is useful for presentation purposes and to help identifying communes with lower living standards (inclusive of contributions due to lower or higher values of predictors).

SMALL AREA ESTIMATES: WHOLE COUNTRY



Bibliography

[1] Australian Bureau of Statistics (2005) *A Guide to Small Area Estimation* http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocument

[2] Goldstein, H. (1989) Restricted unbiased iterative generalised least squares estimation. Biometrika, 76, 622-23.

[3] Goldstein, H. (1995) *Multilevel statistical models. London* [Online]. Available at: http://www.ats.ucla.edu/stat/examples/msm_goldstein/goldstein.pdf

[4] Haughton, D. & Nguyen, P. (2008) Multilevel models and inequality in Vietnam, to appear, *Journal of Data Science*.

[5] Moura, F.A.S. &, Holt, D. (1999) Small area estimation using multilevel models. *Survey Methodology*, 25(1), p.73-80.

[6] Moura, F.A.S (1994) Small area estimation using multilevel models. PhD thesis. University of Southampton.